EMOTION PRESERVATION FOR ONE-SHOT SPEAKER ANONYMIZATION USING MCADAMS

Yamini Sinha[1], *Andreas Wendemuth,*[2], *Ingo Siegert*[1]

[1]*Mobile Dialog Systems, Institute for Information Technology and Communications,*
Otto von Guericke University Magdeburg, Germany
[2]*Cognitive Systems Group, Institute for Information Technology and Communications,*
Otto von Guericke University Magdeburg, Germany
yamini.sinha@ovgu.de, andreas.wendemuth@ovgu.de, ingo.siegert@ovgu.de

Abstract: With the widespread use of voice-assistant, a lot of effort has been made by researchers to improve the usage of these systems. Especially speech recognition and speech understanding are in the focus of research but also the analyses of prosodic information, e.g. emotion recognition is just around the corner.

Similarly, a growing concern for (speech) data privacy arises. Speech data consists of information that can be used to identify a speaker, such as gender, emotional state, and thus a stronger need for protection against misuse is arising. Together with the requirement of easy-to-use anonymization that fits into the easy use of existing voice assistants, the current paper analyses the McAdams anonymization technique, as an approach that can be used without any pre-training. Using a highly expressive German speech database, the performance regarding anonymity, automatic speech recognition performance, and emotion preservation for different parameter ranges of McAdams are analyzed.

1 Introduction

In recent years, the market for commercial voice assistants has been continuously rising [1]. While there is an increase in the popularity of voice assistants in daily usage, the user input data (speech data) is stored and processed on a cloud platform, which raises the concern of data privacy for many. In a 2019 Voice report by Microsoft, 41% of users report concerns around trust, privacy, and passive listening [2]. The skepticism regarding the collected user speech data, which is sensitive and private, for other unauthorized usage is a barrier for adaptation of voice assistants in public and healthcare interactions [3].

For a safer and more convenient integration of future voice assistants, data security requirements cannot be compromised. Speaker anonymization is one of the solutions for such a scenario, which transforms the original user voice to become unidentifiable. But anonymization algorithms are often complex and require user-dependent training, which is hindering the ease of use of voice assistants. Therefore, one possibility would be to use a one-shot speaker anonymization approach that does not require any training, e.g. McAdams coefficient transformation [4]. This technique is simply based on speech processing techniques. The challenge of such an algorithm is to accurately preserve the speech content and prosodic information while removing speaker information. For speech consisting of emotions like sadness, anger, happiness, etc., anonymization may alter the emotional information. The contribution of this paper is the first analysis of the emotion preservation ability of the McAdams anonymization algorithm. Especially due to its use without the training phase the McAdams anonymization technique is particularly suitable for voice assistants in public spaces [5].

2 Related Work

An initiative to develop anonymization solutions for speech technology was provided by the VoicePrivacy Challenge, conducted in 2020 [6]. The task at hand was to hide speakers' identities by transforming the speaker's voice to sound like another speaker, not corresponding to any real speaker. To do so, two baselines were established: 1) anonymization using x-vectors, and 2) anonymization using McAdams coefficient.

Another similar method with the aim to transform original speech without losing any linguistic information is Vocal Tract Length Normalization (VTLN) [7]. In VTLN-based voice conversion, the original voice of the speaker is modified to identify the voice as of the target speaker by warping the frequency axis of the amplitude spectrum of the speaker's voice for the purpose of normalization. The results in [7] showed that piece-wise warping with several parameters showed better precision in warping. However, increasing the number of parameters for warping the source and target spectra were overfitting which interrupted the naturalness of the output speech signal. A similar robotic-sounding output speech is observed from McAdams generated speech. However, to obtain a more intelligible and naturalistic sounding artificial speech, a more complex model might be required. Additionally, in the VPC baseline experiments, the analysis concentrated on the preservation of speech context and intelligibility instead of emotion preservation. Therefore, insights into the performance of privacy preservation while preserving the emotional content of the speech are an interesting aspect to be obtained. One of the very few papers investigating emotion preservation for speaker conversion is [8], unlike in the current paper, [8] rely on training of voice conversion models.

3 Methods

Dataset : As speech data, we rely on the high-quality recordings of the Emotional DataBase (EmoDB) [9]. It contains about 494 utterances with seven emotions: anger, boredom, fear, disgust, happiness, neutral, and sadness, recorded by 10 German speakers (5 female and 5 male). The data was recorded at a 48 kHz sampling rate and then down-sampled to 16 kHz. This database is used as a benchmark in many different experiments related to speech synthesis, emotion recognition, or acoustic analyses.

McAdams Anonymization: A one-shot anonymization technique using the McAdams coefficient is used for speaker anonymization. This is achieved by applying a shift to the formant positions in a speech utterance, thereby adjusting the timbre or spectral envelope [4]. The degree of formant manipulation, performed at the frame level, is controlled by the McAdams coefficient (α). The speech frame is analyzed using Linear Predictive Coding (LPC) to extract the source and the filter coefficients. While the source is set aside for re-synthesis, the filter coefficients are used to determine the shift in pole positions (determined by α) which further alters the speech features. A new set of poles, which includes shifted imaginary term and unaltered real term of the original poles, is converted back to LPC coefficients. The residual and the new LPC coefficients are combined to resynthesize the new anonymized speech frame in the time domain. A detailed explanation is given in [4]. This approach of speaker anonymization requires no training or large amounts of training data. It simply alters the original speech signal using signal processing techniques to change the voice impressions of the speaker.

In this experiment, EmoDB speech data is transformed using McAdams in an attempt to anonymize speakers. Anonymization of this dataset is performed at varying degrees by changing the McAdams coefficient from 0.5 to 1.0 and the filter length from 15 to 25. This results in a total of 59,774 anonymized utterances.

Evaluation: In order to effectively assess the emotion preservation for anonymized speakers' speech while examining the intelligibility, we evaluated three aspects: 1) the success of anonymity of the EmoDB-McAdams generated speech, 2) the ASR-performance of the new transformed speech, and finally, 3) the emotion recognition ability of anonymized samples.

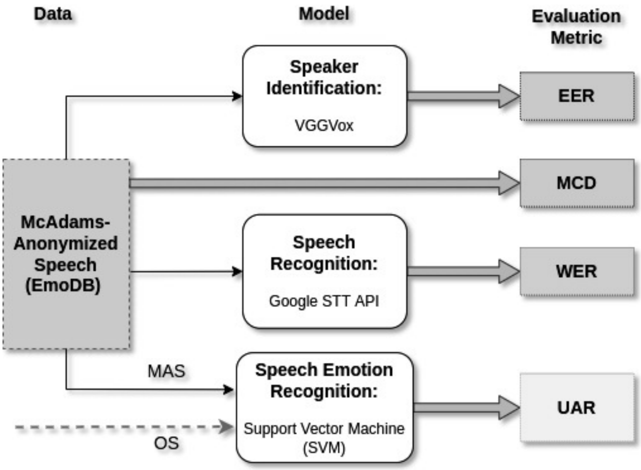


Figure 1 – Utilized evaluation plan for the assessment of anonymity, ASR-performance, and emotion preservation. OS is the original EmoDB speech and MAS is anonymized speech of EmoDB using McAdams coefficient.

- **Anonymity:** A pre-trained speaker recognition model was used to identify the speaker from the anonymized speech. VGGVox, developed by [10], is based on a VGG-M architecture [11]. It adapts a deep-CNN architecture, in contrast to traditional methods that require hand-crafted features, allowing minimal pre-processing of audio data. This model was trained on a large-scale dataset called VoxCeleb1, consisting of over 140,000 utterances by 1,251 celebrities with a wide range of different ages, accents, ethnicity, etc. Ergo, the model learns speaker-specific cues and prosody mannerisms comprehensively, including emotional characteristics [12].
  The original (unaltered) speech samples were used as enrollment data and the anonymized (altered using McAdams) speech was used as test data. A Euclidean distance (or a cosine distance) is then used to compare feature vectors of the test speech with the enroll speech samples. Finally, speaker identification correctness is true (given value: 1) when the predicted speaker id is the same as the test speaker id. In the case of speaker privacy, higher instances of misrecognition would be desirable. Additionally, we also evaluate the spectral differences between altered and unaltered speech samples using Mel-cepstral Distortion (MCD), see section Section 4.
- **ASR-performance:** Upon transforming the original speech to an anonymized one, which possibly degrades the quality of the audio and hence, the intelligibility[1] of the resulting speech may decline. We also assume that the ASR-performance as an aspect of the intelligibility may also differ for different degrees of speaker anonymity with varying α values. To evaluate this, we used the widely known Google Cloud Speech-to-Text engine

[1] It has to be noted that we opt for a more objective measure regarding the intelligibility, as our research actually is aimed to be used for the anonymization of interaction with technical systems.

(STT)[2]. In previous experiments, this engine has proven to be best suited for different languages, background noise, and spontaneous speech [13, 14].

- **Emotion Preservation:** For feature extraction, we used the "emobase" feature set provided by openSMILE [15]. It comprises 988 features derived from 19 functionals calculated for 54 Low Level Descriptors. Afterwards, we normalized (standardization) the values to eliminate differences between the data samples [16]. As a recognition system, an SVM with a linear kernel and a cost factor of 1 was utilized with WEKA [17]. This setup of Speech Emotion Recognition (SER) has proven to generally achieve high recognition performances [18, 19]. As a validation scheme, we applied a Leave-one-speaker-out validation.
  Furthermore, two different experiments were conducted to determine the emotion label using either the Original Speech (OS) or the McAdams-Anonymized Speech (MAS). In experiment 1, the SER is trained and tested using MAS with the identical α-values (within) and in experiment 2 the model is trained with OS and tested against MAS over all α values individually (cross).

4 Measures

Mel-cepstral distortion (MCD) is an objective evaluation, widely used in assessing synthetic speech in voice conversion. It measures the differences between two sequences of mel cepstra, by comparing their spectral distances [20]. For a successful voice conversion, a converted speech must be as similar to the target speech and as different from the source speech as possible, and hence, lower MCD values (when comparing converted speech and target speech) are expected. Whereas, in the case of speaker anonymization using a one-shot method, higher differences (higher MCD value) indicate that the anonymized speech is more different from the original speech. In order to calculate MCD, we calculate the root of the sum of the squared difference between mel-cepstral coefficients (MCEP) of the two speech, as shown in Equation 1:

$$MCD[DB]=\frac{10}{ln10}\sqrt{2\Sigma_{i=1}^{j}(m_{k,i}^{t}-m_{k,i}^{c})^2} \quad (1)$$

Where $\hat{y} = m_{k,i}^{t}$ and $y = m_{k,i}^{c}$ are the two speech samples at frame k, while i denotes the ith coefficient of all the j coefficients of each MCEP vector. With decreasing α, a higher MCD value and degradation in the ASR-performance is expected. The preservation of linguistic information, or lack thereof, can be measured using WER.

Word error rate (WER) compares a hypothesized text (A) to a reference text (B) and evaluates the minimum edit distance (Levenshtein distance) by counting the number of deleted, substituted, and inserted words in text A. It is calculated by adding all the misrecognized words over the total number of words in text B. We used JiWER[3] library supported by python.

Unweighted average recall (UAR) is a common performance metric for emotion recognition performance [21]. It is called over each validation step over all emotion classes available for one speaker. Finally, the UAR and UAP were calculated as the average over all speakers.

5 Results

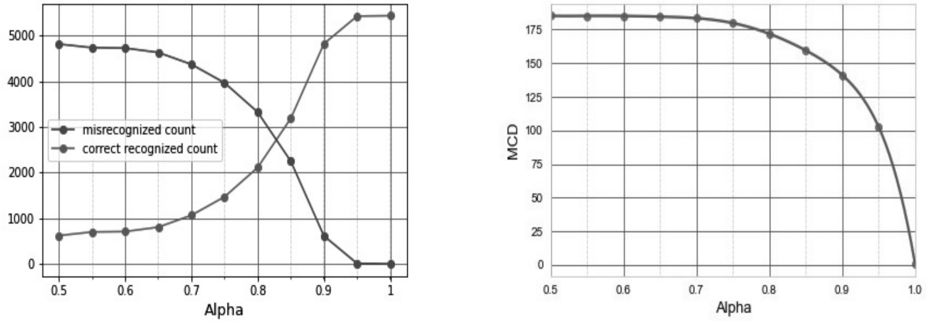
The experimental results are threefold: 1) the success of speaker anonymization using one-shot anonymization technique, 2) intelligibility of the anonymized (altered) speech, and finally 3) the

[2] https://cloud.google.com/speech-to-text/docs.

[3] https://pypi.org/project/jiwer/

emotions preserved. All results presented in this paper are averaged over all speech samples and filter lengths for each α value.

Speaker Anonymization: For the speaker recognition task, a pre-trained deep-convolutional neural network called VGGVox is used. The original speech was used as enrollment data and the EmoDB-anonymized speech was used as the test data. The predicted speaker-id is obtained using a similarity metric called euclidean distance. As the α value gets closer to 0.5, higher anonymization is achieved [4]. As seen in the Figure 2a, for α value closer to 0.5, a higher number of misrecognition events occur, whereas, at α = 1, the altered speech is very close to the original speech, and therefore, the recognition rate is higher.



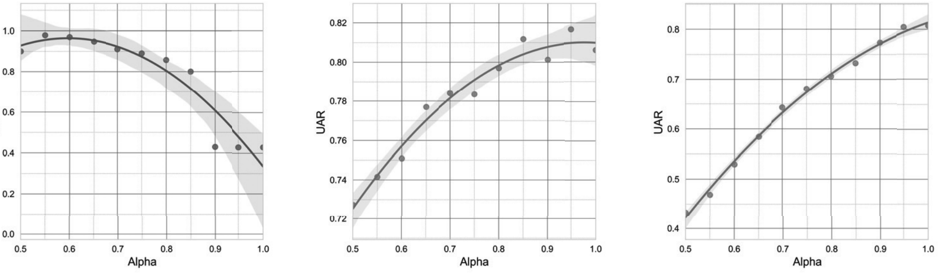
(a) Misrecognized versus correctly recognized speaker samples. **(b)** MCD values, with regression fitting.

Figure 2 – Evaluation of anonymization metrics for different α-based anonymized speech samples of EmoDB. The total number of files is 59,774, higher MCD values imply a higher spectral difference.

A linear relation between alpha values and the success of speaker anonymization can be established. This is also supported by the correlation between α and MCD values, seen in Figure 2b. Lower the alpha value higher is the MCD, approximately 185 dB, and 0.7 dB at α = 1. Since a higher alpha value results in a lower MCD value, anonymization does not take place, making it easier to recognize the speaker.

ASR-Performance: To evaluate the degradation of intelligibility in the EmoDB-anonymized speech, we used an ASR to generate transcription and further, evaluated the WER. Figure 3a follows a similar trend as the results from speaker anonymization. For α values closer to 1, WER is as low as 40%, which is still quite high. We observed 100% WER at α = 0.5, confirming that higher degrees of anonymization using the McAdams algorithm is accompanied by degradation in (technical) intelligibility.

Emotion Preservation: Another interesting aspect is to validate the emotion preservation from the one-shot anonymization technique using McAdams. An SER is used to determine the emotion label in the Original Speech (OS) and the McAdams-Anonymized Speech (MAS) in two phases. In phase 1, the SER is trained and tested using MAS with similar parameters (within) and in phase 2 the model is trained with OS and tested against MAS with specific parameter settings (cross). Both phases are repeated for all intended filter orders and α-values.The results indicate that emotions are still preserved after anonymization with approximately 65% UAR and 78% UAR for cross and within experiments respectively. Figures 3b and 3c show the relation between increasing α value and UAR, which indicates that emotional content degrades with increasing degree of anonymity.



(a) ASR performance (lower WER is better) **(b)** Within Emotion recognition performance **(c)** Cross Emotion recognition performance

Figure 3 – Relation of ASR performance (left) and emotion recognition performance (middle, right) with changing α-values.

6 Outlook and Conclusion

In this paper, we used a one-shot anonymization technique using McAdams coefficient to alter the speaker's voice such that the identity can be hidden. To do so, we used an emotional speech database called EmoDB. We evaluated the preservation of emotional content in altered speech and degradation of ASR-performance with an increasing degree of anonymization. We observed a decline in WER with increasing α value, as the anonymized speech is similar sounding to that of the original speech with α closing 1. Furthermore, the trend of degradation in intelligibility observed is inline with conclusions in [4].

In terms of emotion preservation, we used a standard SVM-based speech emotion recognition system. When the SER was trained on original speech and tested on anonymized speech, we observed an absolute drop of 14% UAR when compared to a system tested on original speech. On the other hand, we observed only a small drop in UAR when SER system is trained and tested on anonymized speech, approximately 1% at best. However, results indicate that emotions are still preserved after anonymization with approximately 65% UAR absolute. Results on all the three aspects: anonymity, degradation of intelligibility and emotion preservation, show a clear relation with the varying α values. Regarding an optimal operation range when using McAdams, we recommend an α value of 0.85 to 0.9. For this range, the WER and SER performance drop is acceptable while the anonymity is still preserved for a one-shot application.

Acknowledgements

This research has been funded by the Federal Ministry of Education and Research of Germany in the project Emonymous (project number S21060A).

References

[1] Kleinberg, S.: *5 ways voice assistance is shaping consumer behavior.* think with Google, 2018. URL `https://perma.cc/U2Y2-Q4WN`. [Online; posted Jan-2018].

[2] Olson, C. and K. Kemery: *Voice report.* 2019. URL `https://advertiseonbing-blob.azureedge.net/blob/bingads/media/insight/whitepapers/2019/04%20apr/voice-report/bingads_2019_voicereport.pdf`.

[3] Wienrich, C., C. Reitelbach, and A. Carolus: *The trustworthiness of voice assistants in the context of healthcare investigating the effect of perceived expertise on the trustworthiness of*

voice assistants, providers, data receivers, and automatic speech recognition. Frontiers in Computer Science, 3, 2021. doi:10.3389/fcomp.2021.685250.

[4] Patino, J., N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans: *Speaker anonymisation using the mcadams coefficient. arXiv preprint arXiv:2011.01130*, 2020.

[5] Siegert, I.: *Speaker anonymization solution for public voice-assistant interactions – presentation of a work in progress development.* In *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, pp. 80–82. 2021.

[6] Tomashenko, N., B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé et al.: *Introducing the voiceprivacy initiative. arXiv preprint arXiv:2005.01387*, 2020.

[7] Sundermann, D. and H. Ney: *Vtln-based voice conversion.* In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No.03EX795)*, pp. 556–559. 2003. doi:10.1109/ISSPIT.2003.1341181.

[8] Nourtel, H., P. Champion, D. Jouvet, A. Larcher, and M. Tahon: *Evaluation of speaker anonymization on emotional speech.* In *1st ISCA Symposium on Security and Privacy in Speech Communication*. 2021.

[9] Burkhardt, F., A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss et al.: *A database of german emotional speech.* In *Interspeech*, vol. 5, pp. 1517–1520. 2005.

[10] Nagrani, A., J. S. Chung, and A. Zisserman: *Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612*, 2017.

[11] Chatfield, K., K. Simonyan, A. Vedaldi, and A. Zisserman: *Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531*, 2014.

[12] Assunção, G. M. S.: *Human Emotion Recognition Through Speech Analysis on Convolutional Neural Networks.* Ph.D. thesis, Universidade de Coimbra, 2019.

[13] Siegert, I., Y. Sinha, O. Jokisch, and A. Wendemuth: *Recognition Performance of Selected Speech Recognition APIs – A Longitudinal Study*, pp. 520–529. Springer, Cham, 2020.

[14] Silber-Varod, V., I. Siegert, O. Jokisch, Y. Sinha, and N. Geri: *A cross-language study of selected speech recognition systems. The Online Journal of Applied Knowledge Management: OJAKM*, 9, pp. 1 – 15, 2021. doi:https://doi.org/10.36965/OJAKM.2021.9(1)1-15.

[15] Eyben, F., M. Wöllmer, and B. Schuller: *openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor.* In *Proc. of the ACM MM-2010*, p. s.p. Firenze, Italy, 2010.

[16] Böck, R., O. Egorow, I. Siegert, and A. Wendemuth: *Comparative Study on Normalisation in Emotion Recognition from Speech*, pp. 189–201. Springer International Publishing, Cham, 2017.

[17] Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten: *The weka data mining software: An update. SIGKDD Explor. Newsl.*, 11(1), pp. 10–18, 2009.

[18] Siegert, I., N. Weisskirchen, J. Krüger, O. Akhtiamov, and A. Wendemuth: *Admitting the addressee detection faultiness of voice assistants to improve the activation performance using a continuous learning framework. Cognitive Systems Research*, 70, pp. 65–79, 2021. doi:10.1016/j.cogsys.2021.07.005.

[19] Siegert, I., A. Lotz, O. Egorow, and S. Wolff: *Utilizing psychoacoustic modeling to improve speech-based emotion recognition*, pp. 625–635. Springer International Publishing, Cham, 2018.

[20] KUBICHEK, R.: *Mel-cepstral distance measure for objective speech quality assessment*. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, pp. 125–128. IEEE, 1993.

[21] SCHULLER, B., A. BATLINER, S. STEIDL, and D. SEPPI: *Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge*. *Speech Commun*, 53, pp. 1062–1087, 2011.